Using Rough Sets, Neural Networks, and Logistic Regression to Predict Compliance with Cholesterol Guidelines Goals in Patients with Coronary Artery Disease

Anil K. Dubey, M.D., M.S.
Laboratory of Computer Science, Massachusetts General Hospital
Boston, Massachusetts.

BACKGROUND

Coronary artery disease is a leading cause of death and disability in the United States and throughout the developed world. Results from large randomized, blinded, placebo-controlled trials have demonstrated clearly the benefit of lowering LDL cholesterol in lowering the risk for coronary artery disease. Unfortunately, despite the quantity of evidence, and the availability of medications that can efficiently lower LDL cholesterol with few side effects, not everyone who could benefit from cholesterol lowering interventions actually receives them. Despite the dissemination of national care guidelines for the evaluation and treatment of cholesterol levels (NCEP – National Cholesterol Education Program), compliance with such guidelines is suboptimal. There clearly is room for improvement in narrowing the gap between evidence based guidelines and actual clinical practice. The ability to classify those patients who are or will likely to be noncompliant on the basis of patient data routinely collected during patient care could be potentially useful by enabling the focusing of limited health care resources to those who are or will be at high risk of being under treated. In order to explore this possibility further, we attempted to create such classifiers of cholesterol guideline compliance. To do this, we obtained data from an ambulatory electronic medical record system at use at the MGH adult primary care practices for over 20 years. We obtained the data from this hierarchically-structured EMR using its own native query language, called MQL (Medical Query Language). Next, we applied to the collected data the machine learning techniques of rough set theory, neural networks (feed forward backpropagation nets), and logistic regression. We did this by using commonly available software that for the most part is freely available via the internet. We then compared the accuracy of the classifier models using the receiver operating characteristic (ROC) area and C-index summary metrics.

METHODS

Definitions of outcome variable and target patient population

For our work we defined the outcome variable of interest to be whether or not a patient is in compliance with established cholesterol guideline standards. The logic of this definition is enumerated in the following decision table:

Is LDL above drug treatment threshold?	Is patient on cholesterol lowering drug therapy?	Compliance with established cholesterol treatment standards
No	**No**	**Yes**
No	**Yes**	**Yes**
Yes	**No**	**No**
Yes	**Yes**	**Yes**

Whether or not someone is on cholesterol lowering therapy was determined by looking for the presence of specific medications: (statins, resins, niacin/nicotinic acid, fibric acid derivates (gemfibrozil, clofibrate). Unfortunately, in some patient populations there is no complete consensus on the exact drug treatment threshold. To increase simplicity and reduce uncertainty, we chose as our target patient population the one whose drug treatment threshold had the greatest consensus: patients with clinically active coronary artery disease. Thus, for the purpose of this work, and for the remainder of this article, unless it is said to the contrary, the implied patient population is the set of patients with CAD.

For individuals with CAD, the ideal LDL goal is less than 100mg/dl. The level at which treatment with pharmacotherapy in this population is categorically recommended is any LDL level greater than or equal to 130mg/dl. In selecting our threshold for use in this work, we decided to use the drug treatment threshold of 130mg/dl – again because we felt this was a drug treatment threshold that carried the greatest current consensus, and thus the least amount of disagreement.

Determination and selection of predictor variables

The variables selected to predict compliance were chosen on the basis of availability from COSTAR and an ad-hoc determination of clinical plausibility. They included the following: **DEMOGRAPHICS** (age, gender, marital Status); **MEDICAL PROBLEMS** (presence /absence of diabetes mellitus, hypertension, stroke, congestive heart failure, peripheral vascular disease, psychiatric diagnosis); **ENCOUNTER DATA** (date of last visit).

Model generation and synthesis

In our analysis we used models based on rough sets, backpropagation feedforward neural networks, and logistic regression. The Rosetta Software package was used to develop the rough set models. NevProp4 was used to develop the neural network models. STATA was used to develop the logistic regression models.

RESULTS

Our preliminary results reveal similar predictive performance among the three different types of models developed thus far, with the neural network model yielding the best predictive performance. Of the best models developed, the rough set model produced an ROC area of 0.613, the neural network model a C-Index of 0.659, and the logistic regression model an ROC area of 0.641.